the sequence 5'-AGACCTTGC-3' and it is suspected that the sample has a possible mutation at the underlined base position, which is the unknown base that will be called by the reference method. The "mutation" probes for the sample sequence may be as follows: 3'-GAAA, 3'-GCAA, 3'-GGAA, and 3'-GTAA, where 3'-GGAA is the wild-type probe.

Suppose further that a reference sequence, which differs from the chip wild-type by one base mutation, has the sequence 5'-AGACATTGC-3' where the mutation base is underlined. The "mutation" probes for the reference sequence may be as follows: 3'-TGAAA, 3'-TGCAA, 3'-TGGAA, and 3'-TGTAA, where 3'-TGTAA is the reference wild-type probe since the reference sequence is known. Although generally the sample and reference sequences were tiled with the same chip wild-type, this is not required, and the tiling methods do not have to be identical as shown by the use of two probe lengths in the example. Thus, the unknown base will be called by comparing the "mutation" probes of the sample sequence to the "mutation" probes of the reference sequence. As before, because each mutation probe is identifiable by the mutation base, the mutation probes' intensities will be referred to as the "base intensities" of their respective mutation bases.

As a simple example of one implementation of the reference method, suppose a gene of interest (target) has the sequence 5'-AAAACTGAAAA-3' (SEQ. ID NO:4). Suppose a reference sequence has the sequence 5'-AAAAC CGAAAA-3' (SEQ ID NO:5), which differs from the target sequence by the underlined base. The reference sequence is marked and exposed to probes on a chip with the target sequence being the chip wild-type. Suppose further that a sample sequence is suspected to have the same sequence as the target sequence except for a mutation at the underlined base position in 5'-AAAACTGAAAA-3' (SEQ ID NO:4). The sample sequence is also marked and exposed to probes on a chip with the target sequence being the chip wild-type. After hybridization and scanning, the following probe intensities (not actual data) were found for the respective complementary probes:

Reference	Sample
3'-TGAC -> 12	3'-GACT -> 11
3'-TGCC -> 9	3'-GCCT -> 30
3'-TGGC -> 80	3'-GGCT -> 60
3'-TGTC -> 15	3'-GTCT -> 6

Although each fluorescence intensity is from a probe, the probes may be identified by their unique mutation base so the bases may be said to have the following intensities:

Reference	Sample
A -> 12	A -> 11
C -> 9	C -> 30
G -> 80	G -> 60
T -> 15	T -> 6

Thus, base A of the reference sequence will be described as having an intensity of 12, which corresponds to the intensity of the mutation probe with the mutation base A. The reference method will now be described as calling the unknown base in the sample sequence by using these intensities.

FIG. 4A illustrates the high level flow of one implementation of the reference method. For illustration purposes, the reference method is described as filling in the columns (identified by the numbers along the bottom) of the analysis table shown in FIG. 4B. However, the generation of an analysis table is not necessary to practice the method. The analysis table is shown to aid the reader in understanding the method.

At step **402** the four base intensities of the reference and sample sequences are adjusted by subtracting the background or "blank" cell intensity from each base intensity. Each set of "mutation" probes has an associated "blank" cell. Suppose that the reference "blank" cell intensity is 1 and the sample "blank" cell intensity is 2. The base intensities are then background subtracted as follows:

Reference	Sample
A -> 12 - 1 = 11	A -> 11 - 2 = 9
C -> 9 - 1 = 8	C -> 30 - 2 = 28
G -> 80 - 1 = 79	G -> 60 - 2 = 58
T -> 15 - 1 = 14	T -> 6 - 2 = 4

Preferably, if a base intensity is then less than or equal to zero, the base intensity is set equal to a small positive number to prevent division by zero or negative numbers in future calculations.

For identification, the position of each base of interest in the reference and sample sequences is placed in column 1 of the analysis table. Also, since the reference sequence is a known sequence, the base at this position is known and is referred to as the reference wild-type. The reference wild-type is placed in column 2 of the analysis table, which is C for this example.

At step **404** the base intensity associated with the reference wild-type (column 2 of the analysis table) is checked to see if it has sufficient intensity to call the unknown base. In this example, the reference wild-type is C. However, the base intensity associated with the wild-type is the G base intensity, which is 79 in this example. This is because the base intensities actually represent the complementary "mutation" probes. The G base intensity is checked by determining if its intensity is greater than a predetermined background difference cutoff. The background difference cutoff is a number that specifies the intensity the base intensities must be above the background intensity in order to correctly call the unknown base. Thus, the base intensity associated with the reference wild-type must be greater than the background difference cutoff or the unknown base is not callable.

If the background difference cutoff is 5, the base intensity associated with the reference wild-type has sufficient intensity (79>5) so a P (pass) is placed in column 3 of the analysis table as shown at step **406**. Otherwise, at step **407** an F (fail) is placed in column 3 of the analysis table.

At step **408** the ratio of the base intensity associated with the reference wild-type to each of the possible bases are calculated. The ratio of the base intensity associated with the reference wild-type to itself will be 1 and the other ratios will usually be greater than 1. The base intensity associated with the reference wild-type is G so the following ratios are calculated:

G:A→79/11=7.2
G:C→79/8=9.9
G:G→79/79=1.0
G:T→79/14=5.6

These ratios are placed in columns 4 through 7 of the analysis table, respectively.

At step **410** the highest base intensity associated with the sample sequence is checked to see if it has sufficient

intensity to call the unknown base. The highest base intensity is checked by determining if the intensity is greater than the background difference cutoff. Thus, the highest base intensity must be greater than the background difference cutoff or the unknown base is not callable.

Again, if the background difference cutoff is 5, the highest base intensity, which is G in this example, has sufficient intensity (58>5) so a P (pass) is placed in column 8 of the analysis table as shown at step 412. Otherwise, at step 413 an F (fail) is placed in column 8 of the analysis table.

At step 414 the ratios of the highest base intensity of the sample to each of the possible bases are calculated. The ratio of the highest base intensity to itself will be 1 and the other ratios will usually be greater than 1. Thus, the highest base intensity is G so the following ratios are calculated:

G:A→58/9=6.4

G:C→58/28=2.3

G:G→58/58=1.0

G:T→58/4=14.5

These ratios are placed in columns 9 through 12 of the analysis table, respectively.

At step 416 if both the reference and sample sequence probes failed to have sufficient intensity to call the unknown base, meaning there is an 'F' in columns 3 and 8 of the analysis table, the unknown base is assigned the code N (insufficient intensity) as shown at step 418. An 'N' is placed in column 17 of the analysis table. Additionally, a confidence code of 9 is placed in column 18 of the analysis table where the confidence codes have the following meanings:

Code	Meaning
0	Probable reference wild-type
1	Probable mutation
2	Reference sufficient intensity, insufficient intensity in sample suggests possible mutation
3	Borderline differences, unknown base ambiguous
4	Sample sufficient intensity, insufficient intensity in reference to allow comparison
5–8	Currently unassigned
9	Insufficient intensity in reference and sample, no interpretation possible

The confidence codes are useful for indicating to the user the resulting analysis of the reference method.

At step 420 if only the reference sequence probes failed to have sufficient intensity to call the unknown base, meaning there is an 'F' in column 3 and a 'P' in column 8 of the analysis table, the unknown base is assigned the code N (insufficient intensity) as shown at step 422. An 'N' is placed in column 17 and a confidence code of 4 is placed in column 18 of the analysis table.

At step 424 if only the sample sequence probes failed to have sufficient intensity to call the unknown base, meaning there is a 'P' in column 3 and a 'F' in column 8 of the analysis table, the unknown base is assigned the code N (insufficient intensity) as shown at step 426. An 'N' is placed in column 17 and a confidence code of 2 is placed in column 18 of the analysis table.

In this example, both the reference and sample sequence probes have sufficient intensity to call the unknown base. At step 428 the ratios of the reference ratios to the sample ratios for each base type are calculated. Thus, the ratio A:A (column 4 to column 9) is placed in column 13 of the analysis table. The ratio C:C (column 5 to column 10) is placed in column 14 of the analysis table. The ratio G:G (column 6 to column 11) is placed in column 15 of the analysis table. Lastly, the ratio T:T (column 7 to column 12) is placed in column 16 of the analysis table. These ratios are calculated as follows:

A:A→7.2/6.4=1.1

C:C→9.9/2.3=4.3

G:G→1.0/1.0=1.0

T:T→5.6/14.5=0.4

The unknown base is called by comparing these ratios of ratios to two predetermined values as follows.

At step 430 if all the ratios of ratios (columns 13 to 16 of the analysis table) are less than a predetermined lower ratio cutoff, the unknown base is assigned the code of the reference wild-type as shown at step 432. Thus, the code for the reference wild-type (as shown in column 2) would be placed in column 17 and a confidence code of 0 would be placed in column 18 of the analysis table.

At step 434 if all the ratios of ratios are less than a predetermined upper ratio cutoff, the unknown base is assigned an ambiguity code that indicates the unknown base may be any one of the bases that has a complementary ratio of ratios greater than the lower ratio cutoff and less than the upper ratio cutoff as shown at step 436. Thus, if the ratio of ratios for A:A, C:C and G:G are all greater than the lower ratio cutoff and less than the upper ratio cutoff, the unknown base would be assigned the code B (meaning "not A"). This is because [illegible] respective base as follows:

A:A→T

C:C→G

G:G→C

so the unknown base would be called as being either C, G, or T, which is identified by the IUPAC code B. This ambiguity code would be placed in column 17 and a confidence code of 3 would be placed in column 18 of the analysis table.

At step 438 at least one of the ratios of ratios is greater than the upper ratio cutoff and the unknown base is called as the base complementary to the highest ratio of ratios. The code for the base complementary to the highest ratio of ratios would be placed in column 17 and a confidence code of 1 would be placed in column 18 of the analysis table.

Assume for the purposes of this example that the lower ratio cutoff is 1.5 and the upper ratio cutoff is 3. Again, the ratios of ratios are as follows:

A:A→1.1

C:C→4.3

G:G→1.0

T:T→0.4

As all the ratios of ratios are not less than the upper ratio cutoff, the unknown base is called the base complementary to the highest ratio of ratios. The highest ratio of ratios is C:C, which has a complementary base G. Thus, the unknown base is called G which is placed in column 17 and a confidence code of 1 is placed in column 18 of the analysis table.

The example shows how the unknown base in the sample nucleic acid sequence was correctly called as base G. Although the complementary "mutation" probe associated with the base G (3'-GCCT) did not have the highest fluorescence intensity, the unknown base was called as base G because the associated "mutation" probe had the highest ratio increase over the other "mutation" probes.

FIG. 5A illustrates the high level flow of another implementation of the reference method. As in the previous

implementation, this implementation also compares the probe intensities of a reference sequence to the probe intensities of a sample sequence. However, this implementation differs conceptually from the previous implementation in that neighboring probe intensities are also analyzed, resulting in more accurate base calling.

As a simple example of this implementation of the reference method, suppose a reference sequence has a sequence of 5'-AAACCCAATCCACATCA-3' (SEQ ID NO:6) and a sample sequence has a sequence of 5'-AAACCCAGTCCACATCA-3' (SEQ ID NO:7), where the mutant base is underlined. Thus, there is a mutation of A to G. Suppose further that the reference and sample sequences are tiled on chips with the reference sequence being the chip wild-type. This implementation of the reference method will be described as identifying this mutation base.

For illustration purposes, this implementation of the reference method is described as filling in a data table shown in FIG. 5B (SEQ ID NO:28, and SEQ ID NO:29). Although the data table contains more data than is required for this implementation, the portions of the data table that are produced by steps in FIG. 5A are shown with the same reference numerals. The generation of a data table is not necessary, however, and is shown to aid the reader in understanding the method. The mutant base position is at position **241** in the reference and sample sequences, which is shown in bold in the data table.

At step **502** the base intensities of the reference and sample sequences are adjusted by subtracting the background or "blank" cell intensity from each base intensity. Preferably, if a base intensity is then less than or equal to zero, the base intensity is set equal to a small positive number to prevent division by zero or negative numbers. In the data table, data **502A** is the background subtracted base intensities for the reference sequence and data **502B** is the background subtracted base intensities for the sample sequence (also called the "mutant" sequence in the data table).

At step **504** the base intensity associated with the reference wild-type is checked to see if it has sufficient intensity to call the unknown base. In this example, the reference wild-type is base A at position **241**. The base intensity associated with the reference wild-type is identified by a lower case "a" in the left hand column. Thus, the base intensities in the data table are not identified by their complements and the reference wild-type at the mutation position has an intensity of **385**. The reference wild-type intensity of **385** is checked by determining if its intensity is greater than a predetermined background difference cutoff. The background difference cutoff is a number that specifies the intensity the base intensities must be over the background intensity in order to correctly call the unknown base. Thus, the base intensity associated with the reference wild-type must be greater than the background difference cutoff or the unknown base is not callable.

If the base intensity associated with the reference wild-type is not greater than the background difference cutoff, the wild-type sequence would fail to have sufficient intensity as shown at step **506**. Otherwise, at step **508** the wild-type sequence would pass by having sufficient intensity.

At step **510** calculations are performed on the background subtracted base intensities of the reference sequence in order to "normalize" the intensities. Each position in the reference sequence has four background subtracted base intensities associated with it. The ratio of the intensity of each base to the sum of the intensities of the possible bases (all four) is calculated, resulting in four ratios, one for each base as shown in the data table. Thus, the following ratios would be calculated at each position in the reference sequence:

A ratio=A/(A+C+G+T)
C ratio=C/(A+C+G+T)
G ratio=G/(A+C+G+T)
T ratio=T/(A+C+G+T)

At position **241**, A ratio would be the wild-type ratio. These ratios are generally calculated in order to "normalize" the intensity data as the photon counts may vary widely from experiment to experiment. Thus, the ratios provide a way of reconciling the intensity variations across experiments. Preferably, if the photon counts do not vary widely from experiment to experiment, the probe intensities do not need to be "normalized."

At step **512** the highest base intensity associated with the sample sequence is checked to see if it has sufficient intensity to call the unknown base. The intensity is checked by determining if the highest intensity sample base is greater than the background difference cutoff. If the intensity is not greater than the background difference cutoff, the sample sequence fails to have sufficient intensity as shown at step **514**. Otherwise, at step **516** the sample sequence passes by having sufficient intensity.

At step **518** calculations are performed on the background subtracted base intensities of the sample sequence in order to "normalize" the intensities. Each position in the sample sequence has four background subtracted base intensities associated with it. The ratios of the intensity of each base to the sum of the intensities of the possible bases (all four) are calculated, resulting in four ratios, one for each base as shown in the data table.

At step **520** if either the reference or sample sequences failed to have sufficient intensity, the unknown base is assigned the code N (insufficient intensity) as shown at step **522**.

At step **524** the normalized base intensity ratios of the reference sequence are subtracted from the normalized base intensity ratios of the sample sequence. Thus, at each position the following calculations are performed:

A Difference=Sample A Ratio−Reference A Ratio
C Difference=Sample C Ratio−Reference C Ratio
G Difference=Sample G Ratio−Reference G Ratio
T Difference=Sample T Ratio−Reference T Ratio

where the reference and sample ratios are calculated at steps **510** and **518**, respectively. The base differences resulting from these calculations are shown in the data table.

At step **526** each position is checked to see if there is a base difference greater than an upper difference cutoff and a base difference lower than a lower difference cutoff. For example, FIG. **5C** shows a graph the normalized sample base intensities minus the normalized reference base intensities. Suppose that the upper difference cutoff is 0.15 and the lower difference cutoff is −0.15 as shown by the horizontal lines in FIG. **5C**. At the mutation position (labeled with a reference 0), the G difference is 0.28 which is greater than 0.15, the upper difference cutoff. Similarly, the A difference is −0.32 which is less than −0.15, the lower difference cutoff. As there is a base difference above the upper difference cutoff and a base difference below the lower difference cutoff, there may be mutation at this position.

If there is neither a base difference above the upper difference cutoff nor a base difference below the lower difference cutoff, the base at that position is assigned the code of the reference wild-type base as shown at step **528**.

At step **530** the ratio of the highest background subtracted base intensity in the sample to the background subtracted

reference wild-type base intensity is calculated. For example, at the mutation position **241** in the data table, the highest background subtracted base intensity in the sample is 571 (base G). The background subtracted reference wild-type base intensity is 385 (base A). Thus, the ratio of 571:385 is calculated and results in 1.48 as shown in the data table.

At step **532** these ratios are compared to a ratio at a neighboring position. The ratio for the n^{th} position is subtracted from the ratio for the r^{th} position, where r=n+1. For example, at the mutation position **241** in the data table, the ratio at position **242** (which equals 1.02) is subtracted from the ratio at position **241** (which equals 1.48). It has been found that a mutant can be confidently detected by analyzing the difference of these neighboring ratios.

FIG. **5**D shows other graphs of data in the data table. Of particular importance is the graph identified as **532** because this is a graph of the calculations at step **532**. The pattern shown in a box in graph **532** has been found to be characteristic of a mutation. Thus, if this pattern is detected, the base is called as the base (or bases) with a normalized difference greater than the upper difference cutoff as shown at step **536**. For example, the pattern was detected and at step **526** it was shown that base G had a normalized difference of 0.28, which is greater than the upper difference cutoff of 0.15. Therefore, the base at position **241** in the sample sequence is called a base G, which is a mutation from the reference sequence (A to G).

If the pattern is not detected at step **534**, the base at that position is assigned the code of the reference wild-type base as shown at step **538**.

This second implementation of the reference method is preferable in some instances as it takes into account probe intensities of neighboring probes. The first implementation may not have detected the A to G mutation in this example.

The advantage of the reference method is that the correct base can be called even in the presence of significant levels of cross-hybridization, as long as ratios of intensities are fairly consistent from experiment to experiment. In practice, the number of miscalls and ambiguities is significantly reduced, while the number of correct calls is actually increased, making the reference method very useful for identifying candidate mutations. The reference method has also been used to compare the reproducibility of experiments in terms of base calling.

IV. Statistical Method

The statistical method is a method of calling bases in a sample nucleic acid sequence. The statistical method utilizes the statistical variation across experiments to call the bases. Therefore, the statistical method is preferable when data from multiple experiments is available and the data is fairly consistent across the experiments. The method compares the probe intensities of a sample sequence to statistics of probe intensities of a reference sequence in multiple experiments.

For simplicity, the statistical method will be described as being used to identify one unknown base in a sample nucleic acid sequence. In practice, the method is used to identify many or all the bases in a nucleic acid sequence.

The unknown base will be called by comparing the probe intensities of a sample sequence to statistics on probe intensities of a reference sequence in multiple experiments. Generally, the probe intensities of the sample sequence and the reference sequence experiments are from chips having the same chip wild-type. However, the reference sequence may or may not be equal to the chip wild-type, as it may have mutations.

A base at the same position in the reference and sample sequences will be associated with up to four mutation probes and a "blank" cell. As before, because each mutation probe is identifiable by the mutation base, the mutation probes' intensities will be referred to as the "base intensities" of their respective mutation bases.

As a simple example of the statistical method, suppose a gene of interest (target) has the sequence 5'-AAAACTGAAAA-3' (SEQ ID NO:4). Suppose a reference sequence has the sequence 5'-AAAACCGAAAA-3' (SEQ ID NO:5), which differs from the target sequence by the underlined base. Suppose further that a sample sequence is suspected to have the same sequence as the target sequence except for a T base mutation at the underlined base position in 5'-AAAACTGAAAA-3' (SEQ ID NO:4). Suppose that in multiple experiments the reference sequence is marked and exposed to probes on a chip. Suppose further the sample sequence is also marked and exposed to probes on a chip.

The following are complementary "mutation" probes that could be used for a reference experiment and the sample sequence:

Reference	Sample
3'-TGAC	3'-GACT
3'-TGCC	3'-GCCT
3'-TGGC	3'-GGCT
3'-TGTC	3'-GTCT

The "mutation" probes shown for the reference sequence may be from only one experiment, the other experiments may have different "mutation" probes, chip wild-types, tiling methods, and the like. Although each fluorescence intensity is from a probe, since the probes may be identified by their unique mutation bases, the probe intensities may be identified by their respective bases as follows:

Reference	Sample
3'-TGAC -> A	3'-GACT -> A
3'-TGCC -> C	3'-GCCT -> C
3'-TGGC -> G	3'-GGCT -> G
3'-TGTC -> T	3'-GTCT -> T

Thus, base A of the reference sequence will be described as having an intensity which corresponds to the intensity of the mutation probe with the mutation base A. The statistical method will now be described as calling the unknown base in the sample sequence by using this example.

FIG. **6** illustrates the high level flow of the statistical method. At step **602** the four base intensities associated with the sample sequence and each of the multiple reference experiments are adjusted by subtracting the background or "blank" cell intensity from each base intensity. Preferably, if a base intensity is then less than or equal to zero, the base intensity is set equal to a small positive number to prevent division by zero or negative numbers.

At step **604** the intensities of the reference wild-type bases in the multiple experiments are checked to see if they all have sufficient intensity to call the unknown base. The intensities are checked by determining if the intensity of the reference wild-type base of an experiment is greater than a predetermined background difference cutoff. The wild-type probe shown earlier for the reference sequence is 3'-TGGC, and thus the G base intensity is the wild-type base intensity. These steps are analogous to steps in the other two methods described herein.

If the intensity of any one of the reference wild-type bases is not greater than the background difference cutoff, the

wild-type experiments fail to have sufficient intensity as shown at step **606**. Otherwise, at step **608** the wild-type experiments pass by having sufficient intensity.

At step **610** calculations are performed on the background subtracted base intensities of each of the reference experiments in order to "normalize" the intensities. Each reference experiment has four background subtracted base intensities associated with it: one wild-type and three for the other possible bases. In this example, the G base intensity is the wild-type, the A, C, and T base intensities being the "other" intensities. The ratios of the intensity of each base to the sum of the intensities of the possible bases (all four) are calculated, giving one wild-type ratio and three "other" ratios. Thus, the following ratios would be calculated:

A ratio=A/(A+C+G+T)

C ratio=C/(A+C+G+T)

G ratio=G/(A+C+G+T)

T ratio=T/(A+C+G+T)

where G ratio is the wild-type ratio and A, C, and T ratios are the "other" ratios. These four ratios are calculated for each reference experiment. Thus if the number of reference experiments is n, there would be 4n ratios calculated. These ratios are generally calculated in order to "normalize" the intensity data, as the photon counts may vary widely from experiment to experiment. However, if the probe intensities do not vary widely from experiment to experiment, the probe intensities do not need to be "normalized."

At step **612** statistics are prepared for the ratios calculated for each of the reference experiments. As stated before, each reference experiment will be associated with one wild-type ratio and three "other" ratios. The mean and standard deviation are calculated for all the wild-type ratios. The mean and standard deviation are also calculated for each of the other ratios, resulting in three other means and standard deviations for each of the bases that is not the wild-type base. Therefore, the following would be calculated:

Mean and standard deviation of A ratios

Mean and standard deviation of C ratios

Mean and standard deviation of G ratios

Mean and standard deviation of T ratios

where the mean and standard deviation of the G ratios are also known as the wild-type mean and the wild-type standard deviation, respectively. The mean and standard deviation of the A, C, and T means and standard deviations are also known collectively as the "other" means and standard deviations.

Suppose that the preceding calculations produced the following data:

A ratios→mean=0.16 std. dev.=0.003

C ratios→mean=0.03 std. dev.=0.002

G ratios→mean=0.71 std. dev.=0.050

T ratios→mean=0.11 std. dev.=0.004

In one embodiment, the steps up to and including step **612** are performed in a preprocessing stage for the multiple wild-type experiments. The results of the preprocessing stage are stored in a file so that the reference calculations do not have to be repeatedly calculated, improving performance. Microfiche Appendices C and D contain the programming code to perform the preprocessing stage.

At step **614** the highest base intensity associated with the sample sequence is checked to see if it has sufficient intensity to call the unknown base. The intensity is checked by determining if the highest intensity unknown base is greater than the background difference cutoff. If the intensity is not greater than the background difference cutoff, the sample sequence fails to have sufficient intensity as shown at step **616**. Otherwise, at step **618** the sample sequence passes by having sufficient intensity.

At step **620** calculations are performed on the four background subtracted intensities of the sample sequence. The ratios of the background subtracted intensity of each base to the sum of the background subtracted intensities of the possible bases (all four) are calculated, giving four ratios, one for each base. For consistency, the ratio associated with the reference wild-type base is called the wild-type ratio, with there being three "other" ratios. Thus, the following ratios are calculated:

A ratio=A/(A+C+G+T)

C ratio=C/(A+C+G+T)

G ratio=G/(A+C+G+T)

T ratio=T/(A+C+G+T)

where ratio G is the wild-type ratio and ratios A, C, and T are the "other" ratios.

Suppose the background subtracted intensities associated with the sample are as follows:

A→310

C→50

G→26

T→100

Then, the corresponding ratios would be as follows:

A ratio=310/(310+50+26+100)=0.64

C ratio=50/(310+50+26+100)=0.10

G ratio=26/(310+50+26+100)=0.05

T ratio=100/(310+50+26+100)=0.21

At step **622** if either the reference experiments or the sample sequence failed to have sufficient intensity, the unknown base is assigned the code N (insufficient intensity) as shown at step **624**.

At step **626** the wild-type and "other" ratios associated with the sample sequence are compared to statistical expressions. The statistical expressions include four predetermined standard deviation cutoffs, one associated with each base. Thus, there is a standard deviation cutoff for each of the bases A, C, G, and T. The localized standard deviation cutoffs allow the unknown base to be called with higher precision because each standard deviation cutoff can be set to a different value. Suppose the standard deviation cutoffs are set as follows:

A standard deviation cutoff→4

C standard deviation cutoff→2

G standard deviation cutoff→8

T standard deviation cutoff→4

The wild-type base ratio associated with the sample is compared to a corresponding statistical expression:

WT ratio≧WT mean−(WT std. dev.*WT base std. dev. cutoff)

where the WT base std. dev. cutoff is the standard deviation cutoff for the wild-type base. As the wild-type base is G, the above comparison solves to the following:

0.05≧0.71−(0.050*8)

0.05≧0.31

which is not a true expression (0.05 is not greater than 0.31).

Each of the "other" ratios associated with the sample is compared to a corresponding statistical expression:

Other ratio>Other mean+(Other std. dev.*Other base std. dev. cutoff)

where the Other base std. dev. cutoff is the standard deviation cutoff for the particular "other" base. Thus, the above comparison solves to the following three expressions:

A→0.64>0.16+(0.003*4) 0.64>0.17

C→0.10>0.03+(0.002*2) 0.10>0.03

T→0.21>0.11+(0.004*4) 0.21>0.13

which are all true expressions.

At step **628** if only the wild-type ratio of the sample sequence was greater than the statistical expression, the unknown base is assigned the code of the reference wild-type base as shown at step **630**.

At step **632** if one or more of the "other" ratios of the sample sequence were greater than their respective statistical expressions, the unknown base is assigned an ambiguity code that indicates the unknown base may be any one of the complements of these bases, including the reference wild-type. In this example, the "other" ratios for A, C, and T were all greater than their corresponding statistical expression. Thus, the unknown base would be called the complements of these bases, represented by the subset T, G, and A. Thus, the unknown base would be assigned the code D (meaning "not C").

If none of the ratios are greater than their respective statistical expressions, the unknown base is assigned the code X (insufficient discrimination) as shown at step **636**.

The statistical method provides accurate base calling because it utilizes statistical data from multiple reference experiments to call the unknown base. The statistical method has also been used to implement confidence estimates and calling of mixed sequences.

V. Pooling Processing

The present invention provides pooling processing which is a method of processing reference and sample nucleic acid sequences together to reduce variations across individual experiments. In the representative embodiment discussed herein, the reference and sample nucleic acid sequences are labeled with different fluorescent markers emitting light at different wavelengths. However, the nucleic acids may be labeled with other types of markers including distinguishable radioactive markers.

After the reference and sample nucleic acid sequences are labeled with different color fluorescent markers, the labeled reference and sample nucleic acid sequences are then combined and processed together. An apparatus for detecting targets labeled with different markers is provided in U.S. application Ser. No. 08/195,889 and is hereby incorporated by reference for all purposes.

FIG. **7** illustrates the pooling processing of a reference and sample nucleic acid sequence. At step **702** a reference nucleic acid sequence is marked with a fluorescent dye, such as a fluorescein. At step **704** a sample nucleic acid sequence is marked with a dye that, upon excitation, emits light of a different wavelength than that of the fluorescent dye of the reference sequence. For example, the sample nucleic acid sequence may be marked with rhodamine.

At step **706** the labelel reference sequence and the labeled sample sequence are combined. After this step, processing continues in the same manner as for only one labeled sequence. At step **708** the sequences are fragmented. The fragmented nucleic acid sequences are then hybridized on a chip containing probes as shown at step **710**.

At step **712** a scanner generates image files that indicate the locations where the labeled nucleic acids bound to the chip. In general, the scanner generates an image file by focusing excitation light on the hybridized chip and detecting the fluorescent light that is emitted. The marker emitting the fluorescent light can be identified by the wavelength of the light. For example, the fluorescence peak of fluorescein is about 530 nm while that of a typical rhodamine dye is about 580 nm.

The scanner creates an image file for the data associated with each fluorescent marker, indicating the locations where the correspondingly labeled nucleic acid bound to the chip. Based upon an analysis of the fluorescence intensities and locations, it becomes possible to extract information such as the monomer sequence of DNA or RNA.

Pooling processing reduces variations across individual experiments because much of the test environment is common. Although pooling processing has been described as being used to improve the combined processing of reference and sample nucleic acid sequences, the process may also be used for two reference sequences, two sample sequences, or multiple sequences by utilizing multiple distinguishable markers.

VI. Comparative Analysis (VIEWSEQ™)

The present invention provides a method of comparative analysis and visualization of multiple experiments. The method allows the intensity ratio, reference, and statistical methods to be run on multiple datafiles simultaneously. This permits different experimental conditions, sample preparations, and analysis parameters to be compared in terms of their effects on sequence calling. The method also provides verification and editing functions, which are essential to reading sequences, as well as navigation and analysis tools.

FIG. **8** illustrates the main screen and the associated pull down menus for comparative analysis and visualization of multiple experiments (SEQ ID NO:9). The windows shown are from an appropriately programmed Sun Workstation. However, the comparative analysis software may also be implemented on or ported to a personal computer, including IBM PCs and compatibles, or other workstation environments. A window **802** is shown having pull down menus for the following functions: File **804**, Edit **806**, View **808**, Highlight **810**, and Help **812**.

The main section of the window is divided into a reference sequence area **814** and a sample sequence area **816**. The reference sequence area is where known sequences are displayed and is divided into a reference name subarea **818** and reference base subarea **820**. The reference name subarea is shown with the filenames that contain the reference sequences. The chip wild-type is identified by the filename with the extension ".wt#" where the # indicates a unit on the chip. The reference base subarea contains the bases of the reference sequences. A capital C **822** is displayed to the right of the reference sequence that is the chip wild-type for the current analysis. Although the chip wild-type sequence has associated fluorescence intensities, the other reference sequences shown below the chip wild-type may be known sequences that have not been tiled on the chip. These may or may not have associated fluorescence intensities. The reference sequences other than the chip wild-type are used for sequence comparisons and may be in the form of simple ASCII text files.

Sample sequence area **816** is where sample or unknown experimental sequences are displayed for comparison with the reference sequences. The sample sequence area is divided into a sample name subarea **824** and sample base subarea **826**. The sample name subarea is shown with filenames that contain the sample sequences. The filename

extensions indicate the method used to call the sample sequence where ".cq#" denotes the intensity ratio method, ".rq#" denotes the reference method, and ".sq#" denotes the statistical method (# indicates the unit on the chip). The sample base subarea contains the bases of the sample sequences. The bases of the sample sequences are identified by the codes previously set forth which, for the most part, conform to the IUPAC standard.

Window **802** also contains a message panel **828**. When the user selects a base with an input device in the reference or sample base subarea, the base becomes highlighted and the pathname of the file containing the base is displayed in the message panel. The base's position in the nucleic acid sequence is also displayed in the message panel.

In pull down menu File **804**, the user is able to load files of experimental sequences that have been tiled and scanned on a chip. There is a chip wild-type associated with each experimental sequence. The chip wild-type associated with the first experimental sequence loaded is read and shown as the chip wild-type in reference sequence area **814**. The user is also able to load files of known nucleic acid sequences as reference sequences for comparison purposes. As before, these known reference sequences may or may not have associated probe intensity data. Additionally, in this menu the user is able to save sequences that are selected on the screen into a project file that can be loaded in at a later time. The project file also contains any linkage of the sequences, where sequences are linked for comparison purposes. Sequences to be saved, both reference and sample, are chosen by selecting the sequence filename with an input device in the reference or sample name subareas.

In pull down menu Edit **806**, the user is able to link together sequences in the reference and sample sequence areas. After the user has selected one reference and one or more sample sequences, the sample sequences can be linked to the reference sequence by selecting an entry in the pull down menu. Once the sequences are linked, a link number **830** is displayed next to each of the linked sequences. Each group of linked sequences is associated with a unique link number, so the user can easily identify which sequences are linked together. Linking sequences permits the user to more easily compare sequences of related interest. The user is also able to remove and display links in this menu.

In pull down menu View 808, the user is able to display intensity graphs for selected bases. Once a base is selected in the reference or-sample base subareas, the user may request an intensity graph showing the hybridized probe intensities of the selected base and a delineated neighborhood of bases near the selected base. Intensity graphs may be displayed for one or multiple selected bases. The user is also able to prepare comment files and reports from this menu.

FIG. **9** illustrates an intensity graph window for a selected base at position **120** (SEQ ID NO:31). The filename containing the sequence data is displayed at **904**. The graph shows the intensities for each of the hybridized probes associated with a base. Each grouping of four vertical bars on the graph, which are labeled as "a", "ic", "g", and "t" on line **906**, shows the background subtracted intensities of probes having the indicated substitution base. In one embodiment, the called bases are shown in red. The wild-type base is shown at line **908**, the called base is shown at line **910**, and the base position is shown at line **912**. In FIG. **9**, the base selected is at position **120**, as shown by arrow **914**. The wild-type base at this position is T; however, the called base is M which means the base is either A or C (amino). The user is able to use intensity graphs to visually compare the intensities of each of the possible calls.

FIG. **10** illustrates multiple intensity graph windows for selected bases (SEQ ID NO:33, SEQ ID NO:34 and SEQ ID NO:35). There are three intensity graph windows **1002**, **1004**, and **1006** as shown. Each window may be associated with a different experiment, where the sequence analyzed in the experiment may be either a reference (if it has associated probe intensity data as in the chip wild-type) or a sample sequence. The windows are aligned and a rectangular box **1008** shows the selected bases' position in each of the sequences (position **162** in FIG. **10**). The rectangular box aids the user in identifying the selected bases.

Referring again to FIG. **8**, in pull down menu Highlight **810**, the user is able to compare the sequences of references and samples. At least four comparisons are available to the user, including the following: sample sequences to the chip wild-type sequence, sample sequences to any reference sequences, sample sequences to any linked reference sequences, and reference sequences to the chip wild-type sequence. For example, after the user has linked a reference and sample sequence, the user can compare the bases in the linked sequences. Bases in the sample sequence that are different from the reference sequence will then be indicated on the display device to the user (e.g., base is shown in a different color). In another example, the user is able to perform a comparison that will help identify sample sequences. After a sample is linked to multiple reference sequences, each base in the sample sequence that does not match the wild-type sequence is checked to see if it matches one of the linked reference sequences. The bases that match a linked reference sequence will then be indicated on the display device to the user. The user may then more easily identify the sample sequence as being one of the reference sequences.

In pull down menu Help **812**, the user is able to get information and instructions regarding the comparative analysis program, the calling methods, and the IUPAC definitions used in the program.

FIG. **11** illustrates the intensity ratio method correctly calling a mutation in solutions with varying concentrations (SEQ ID NO:10, SEQ ID NO:11, SEQ ID NO:12, SEQ ID NO:13, SEQ ID NO:14, SEQ ID NO:15, SEQ ID NO:16, SEQ ID NO:17, and SEQ ID NO:18). A window **1102** is shown with a chip wild-type **1104** and a mutant sequence **1106**. The mutant sequence differs from the chip wild-type at the position indicated by the rectangular box **1108**. The chip wild-type and mutant sequences are a region of HIV Pol Gene spanning mutations occurring in AZT drug therapy.

There are seven sample sequences that are called using the intensity ratio method. The sample sequences are actually solutions of different proportions of the chip wild-type sequence and the mutant sequence. Thus, there are sample solutions **1110**, **1112**, **1114**, **1116**, **1118**, **1120**, and **1122**. The solutions are 15-mer tilings across the chip wild-type with increased percentages of the mutant sequence from 0 to 100% by weight. The following shows the proportions of the sample solutions:

Sample Solution	Chip Wild-Type:Mutant
1110	100:0
1112	90:10
1114	75:25
1116	50:50
1118	25:75
1120	10:90
1122	0:100

For example, sample solution **1114** contains 75% chip wild-type sequence and 25% mutant sequence.

Now referring to the bases called in rectangular box **1108** for the sample solutions, the intensity ratio method correctly calls sample solution **1110** as having a base A as in the chip-wild type sequence. This is correct because sample solution **1110** is 100% chip wild-type sequence. The intensity ratio method also calls sample solution **1112** as having a base A because the sample solution is 90% chip wild-type sequence.

The intensity ratio method calls the identified base in sample solutions **1114** and **1116** as being an R, which is an ambiguity IUPAC code denoting A or G (purine). This also a correct base call because the sample solutions have from 75% to 50% chip-wild type sequence and from 25% to 50% mutation sequence. Thus, the intensity ratio method correctly calls the base in this transition state.

Sample solutions **1118**, **1120**, and **1122** are called by the intensity ratio method as having a mutation base G at the specified location. This is a correct base call because the sample solutions primarily consist of the mutation sequence (75%, 90%, and 100% respectively). Again, the intensity ratio method correctly called the bases.

These experiments also show that the base calling methods of the present invention may also be used for solutions of more than one nucleic acid sequence.

FIG. **12** illustrates the reference method correctly calling a mutant base where the intensity ratio method incorrectly called the mutant base (SEQ ID NO:36, SEQ ID NO:37, SEQ ID NO:38, and SEQ ID NO:39). There are three intensity graph windows **1202**, **1204**, and **1206** as shown. The windows are aligned and a rectangular box **1208** outlines the bases of interest. Window **1202** shows a sample sequence called using the intensity ratio method. However, the base in the rectangular box **1208** was incorrectly called base c, as there is actually a base A at that position. The intensity ratio method incorrectly called the base as C because the probe intensity associated with base C is much higher than the other probe intensities.

Window **1204** shows a reference sequence called using the intensity ratio method. As the reference sequence is known, it is not necessary to know the method used to call the reference sequence. However, it is important to have probe intensities for a reference sequence to use the reference method. The reference sequence is called base C at the position indicated by the rectangular box.

Window **1206** shows the sample sequence called using the reference method. The reference method correctly calls the specified base as being base A. Thus, for some cases the reference method is preferable to the intensity ratio method because it compares probe intensities of a sample sequence to probe intensities of a reference sequence.

VII. EXAMPLES

Example 1

The intensity ratio method was used in sequence analysis of various polymorphic HIV-1 clones using a protease chip. Single stranded DNA of a 382 nt region was used with 4 different clones (HXB2, SF2, NY5, pPol4mut18). Results were compared to results from an ABI sequencer. The results are illustrated below:

	ABI		Protease Chip	
	Sense	Antisense	Sense	Antisense
No call	0	4	9	4
Ambiguous	6	14	17	8
Wrong call	2	3	3	1
TOTAL	8	21	29	13

SUMMARY

ABI (sense) - 99.5%
Chip (sense) - 98.1%
ABI (antisense) - 98.6%
Chip (antisense) - 99.1%

Example 2

HIV protease genotyping was performed using the described chips and CALLSEQ™ intensity ratio calculations. Samples were evaluated from AIDS patients before and after ddI treatment. Results were confirmed with ABI sequencing.

FIG. **13** illustrates the output of the VIEWSEQ™ program with four pretreatment samples and four posttreatment samples (SEQ ID NO:19, SEQ ID NO:20, SEQ ID NO:21, SEQ ID NO:22, SEQ ID NO 23, SEQ ID NO:24, SEQ ID NO:25, SEQ ID NO:26, and SEQ ID NO:27). Note the base change at position **207** where a mutation has arisen. Even adjacent two additional mutations (gt), the "a" mutation has been properly detected.

VIII. APPENDICES

The Microfiche appendices (copyright Affymetrix, Inc.) provide C++ source code and header files for implementing the present invention. Appendix A contains the source code files (.cc files) for CALLSEQ™, which is a base calling program that implements the intensity ratio, reference, and statistical methods of the present invention. Appendix B contains the header files (.h files) for CALLSEQ™. Appendices C and D contain the source code and header files, respectively, for a program that performs a preprocessing stage for the statistical method of CALLSEQ™.

Appendix E contains the source code and header files for VIEWSEQ™, which is a comparative analysis and visualization program according to the present invention. Appendices A–E are written for a Sun Workstation.

The above description is illustrative and not restrictive. Many variations of the invention will become apparent to those of skill in the art upon review of this disclosure. Merely by way of example, while the invention is illustrated with particular reference to the evaluation of DNA (natural or unnatural), the methods can be used in the analysis from chips with other materials synthesized thereon, such as RNA. The scope of the invention should, therefore, be determined not with reference to the above description, but instead should be determined with reference to the appended claims along with their full scope of equivalents.

SEQUENCE LISTING

(1) GENERAL INFORMATION:

(i i i) NUMBER OF SEQUENCES: 39

(2) INFORMATION FOR SEQ ID NO:1:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 15 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (oligonucleotide)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:1:

ATGTGGACAG TTGTA 15

(2) INFORMATION FOR SEQ ID NO:2:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 15 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (oligonucleotide)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:2:

ATGTGGATAG TTGTA 15

(2) INFORMATION FOR SEQ ID NO:3:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 15 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (oligonucleotide)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:3:

ATGTGGAKAG TTGTA 15

(2) INFORMATION FOR SEQ ID NO:4:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 11 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (oligonucleotide)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:4:

AAAACTGAAA A 11

(2) INFORMATION FOR SEQ ID NO:5:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 11 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (oligonucleotide)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:5:

-continued

AAAACCGAAA A 11

(2) INFORMATION FOR SEQ ID NO:6:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 17 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (oligonucleotide)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:6:

AAACCCAATC CACATCA 17

(2) INFORMATION FOR SEQ ID NO:7:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 17 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (oligonucleotide)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:7:

AAACCCAGTC CACATCA 17

(2) INFORMATION FOR SEQ ID NO:8:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 31 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (oligonucleotide)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:8:

GGGGAAGCAG ATTTGGGTAC CACCCAAGTA T 31

(2) INFORMATION FOR SEQ ID NO:9:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 31 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (oligonucleotide)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:9:

GGGGAAGCAG ATTTGAAMAC CACCCAAGTA T 31

(2) INFORMATION FOR SEQ ID NO:10:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 59 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (oligonucleotide)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:10:

GCATTAGTAG AGATATGTAC AGAAATGGAA AAGGAAGGGA AAATTTCAAA AATTGGGCC 59

(2) INFORMATION FOR SEQ ID NO:11:

-continued

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 59 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (oligonucleotide)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:11:

GCATTAGTAG AAATTTGTAC AGAGATGGAA AAGGAAGGGA AAATTTCAAA AATTGGGCC 59

(2) INFORMATION FOR SEQ ID NO:12:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 59 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (oligonucleotide)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:12:

GCATTAGTAG AGATATGGAG AGRARDGGRA ANNNAAGGGA AAATTNNNAA AATTGGGCC 59

(2) INFORMATION FOR SEQ ID NO:13:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 59 base pairs
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (oligonucleotide)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:13:

GCATTAGTAG AGATATGKAS AGRARDGGRA ANNNAAGGGA AAAKTNNNAA AATTGGGCC 59

(2) INFORMATION FOR SEQ ID NO:14:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 59 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (oligonucleotide)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:14:

GCATTAGTAG AGATATGKAS AGRRRDGGRA ANNNAAGGGA AAADTYNNAA AATTGGGCC 59

(2) INFORMATION FOR SEQ ID NO:15:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 59 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (oligonucleotide)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:15:

GCATTAGTAG AGATATGTAS AGRRADGGAA ANGGAAGGGA AAATTNNNNA AATTGGGCC 59

(2) INFORMATION FOR SEQ ID NO:16:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 59 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (oligonucleotide)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:16:

GCATTAGTAG AGATATGTAC AGRGAGGGAA ANGGAAGGGA AAATTNNNNA AATTGGGCC 59

(2) INFORMATION FOR SEQ ID NO:17:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 59 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (oligonucleotide)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:17:

GCATTAGTAG AGATATGTAS AGRGAGGGAA ANGGAAGGGA AAATTNNNNA AATTGGGCC 59

(2) INFORMATION FOR SEQ ID NO:18:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 59 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (oligonucleotide)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:18:

GCATTAGTAG GAGGNNNGAC AGGGRKGGAA ANNMAAGGGA AAAKTNNNAA AATTGGGCC 59

(2) INFORMATION FOR SEQ ID NO:19:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 160 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (oligonucleotide)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:19:

TCGAGATAAT CTATGTCCTC GTCTACTATG TCATAATCTT CTTTACTTAA ACGGTCCTTT 60

TACCTTTGGT TTTTACTATC CCCCTTAACC TCCAAAATAG TTTCATTCTG TCATGCTAGT 120

CTATGGACAT CTTTAGACAC CTGTATTTCG ATATCCATGT 160

(2) INFORMATION FOR SEQ ID NO:20:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 160 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (oligonucleotide)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:20:

NNGAGATANN NTATGTCCTC GTCYACTATG TNANNNNNNN NNNNNNNNAA ACGGTCCTNN 60

NNNNNNNNNN NNNNNNNNNN CNNCNTAACC TCCAAAATAN NNNNNNTCTN NNNNANNNNT 120

CTANNNGNAG NNNNAGANAR NCCNNNNNNN NNATNCATGT 160

(2) INFORMATION FOR SEQ ID NO:21:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 160 base pairs

-continued

(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (oligonucleotide)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:21:

```
TCGAGATAAT CTATGTCCTC GTCTACTATG TCATAATNNN NNNNACTTAA ACGGTCCTTT      60
TACCTTTGGT TTTTACTATC CCCCTTAACC TCCAAAATAG TTTCATTCTG NCATANNAGT     120
CTATGNGNNG NNNTAGACAG NCCNNNNTCG ATATCCATGT                           160
```

(2) INFORMATION FOR SEQ ID NO:22:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 160 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (oligonucleotide)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:22:

```
TCGAGATAAT CTATGTCCTC GTCTACTATG TCATAATCTT CTTTACTTAA ACGGTCCTTT      60
TACCTTTGGT TTTTACTATC CNNCTTAACC TCCAAAATAG TTTCATTCTG TCATACTAGT     120
CTATGGGTAG CTTTAGACCN CCGTATTTCG ATATCCATGT                           160
```

(2) INFORMATION FOR SEQ ID NO:23:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 160 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (oligonucleotide)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:23:

```
TCGAGATAAT CTATGTCCTC GTCTACTATG TCATAATCTT CTTTACTTAA ACGGTCCTTT      60
TACCTTTGGT TTTTACTATC CCNCTTAACC TCCAAAATAG TTTCATTCTG TCATACTAGT     120
CTATGGGTAG CTTTAGACCC CCGTATTTCG ATATCCATGT                           160
```

(2) INFORMATION FOR SEQ ID NO:24:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 160 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (oligonucleotide)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:24:

```
NCGGGATANT NTATGTCCTC GTCYACTATG TCANNNNNCN NNCNNNNCAA ACGGTCCNCC      60
NNNNNCNNNN NNCNNCYANG AANCYCAACC TCCAAAATAN NNNNNNTCTN NNNNANNNCN     120
CTNNNNNNAG NGNNAGACAC CTGTATNNNN NTATNCAYGT                           160
```

(2) INFORMATION FOR SEQ ID NO:25:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 160 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

-continued

(i i) MOLECULE TYPE: DNA (oligonucleotide)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:25:

```
TCGRGATAAT CTATGTCCTC GTCTACTATG TCATAATCCN NNCNNCTCAA ACGGTCCTYC      60
CNNNNYTGGT TNYTACTATC CCCCTTAACC TCCAAAATAG TTTCATTCTG NCATACNNST     120
CTANNNNNAG NGTTAGACAC CTGTATTTCG ATATCCATGT                           160
```

(2) INFORMATION FOR SEQ ID NO:26:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 160 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (oligonucleotide)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:26:

```
TCGAGATAAT CTATGTCCTC GTCTACTATG TCATAATCCN NCCTACTCAA ACGGTCCTTC      60
TACCTTTGGT TTTTACTATC CMCCTTAACC TCCAAAATAG TTTCATTCTG TCATACTAGT     120
CTATGAGTAG CTTTAGACAC CTGTATTTCG ATATCCATGT                           160
```

(2) INFORMATION FOR SEQ ID NO:27:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 160 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (oligonucleotide)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:27:

```
TCGAGATAAT CTATGTCCTC GTCTACTATG TCATAATCTT CTTTACYCAA ACGGTCCTNC      60
TACCTTTGGT TTTTACTATC CCMCTTAACC TCCAAAATAG TTTCATTCTG TCATACTAGT     120
CTATGAGTAG CTTTAGACAC CTGTATTTCG ATATCCATGT                           160
```

(2) INFORMATION FOR SEQ ID NO:28:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 17 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (oligonucleotide)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:28:

```
AAACCCAATC CACATCM                                                     17
```

(2) INFORMATION FOR SEQ ID NO:29:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 17 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (oligonucleotide)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:29:

```
MMACNCANNC CACANNM                                                     17
```

(2) INFORMATION FOR SEQ ID NO:30:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 11 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (oligonucleotide)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:30:

TTGGGTACCA C 11

(2) INFORMATION FOR SEQ ID NO:31:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 11 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (oligonucleotide)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:31:

TTGAAMACCA C 11

(2) INFORMATION FOR SEQ ID NO:32:

(i) SEQUENCE CHARACTERISTICS:
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (oligonucleotide)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:32:

ACAGAAATGG A 11

(2) INFORMATION FOR SEQ ID NO:33:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 11 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (oligonucleotide)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:33:

AGAGRATDGG R 11

(2) INFORMATION FOR SEQ ID NO:34:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 11 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (oligonucleotide)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:34:

ASAGRRADGG A 11

(2) INFORMATION FOR SEQ ID NO:35:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 11 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single

-continued

(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (oligonucleotide)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:35:

ACAGGGRRGG A 11

(2) INFORMATION FOR SEQ ID NO:36:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 11 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (oligonucleotide)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:36:

CTGGGGGGTA T 11

(2) INFORMATION FOR SEQ ID NO:37:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 11 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (oligonucleotide)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:37:

CTGGCCSGTG T 11

(2) INFORMATION FOR SEQ ID NO:38:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 11 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (oligonucleotide)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:38:

CTGGGCGGTA T 11

(2) INFORMATION FOR SEQ ID NO:39:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 11 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (oligonucleotide)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:39:

CTGGCACGTG T 11

What is claimed is:

1. A computer program product that identifies an unknown base in a sample nucleic acid sequence, comprising:

computer code that receives a plurality of signals corresponding to probe intensities for a plurality of nucleic acid probes, each probe intensity indicating an extent of hybridization of a nucleic acid probe with at least one nucleic acid sequence including said sample sequence, and each nucleic acid probe differing from each other by at least a single base;

computer code that performs a comparison of said plurality of probe intensities to each other;

computer code that generates a base call identifying said unknown base according to results of said comparison and said sequences of said nucleic acid probes; and

a computer readable medium that stores said computer codes.

2. A computer program product that identifies an unknown base in a sample nucleic acid sequence, comprising:

computer code that receives a plurality of signals corresponding to probe intensities for a plurality of nucleic acid probes, each probe intensity indicating an extent of hybridization of a nucleic acid probe with said sample sequence, and each nucleic acid probe differing from each other by at least a single base;

computer code that calculates a ratio of a higher probe intensity to a lower probe intensity;

computer code that generates a base call identifying said unknown base according to a base of a nucleic acid probe having said higher probe intensity if said ratio is greater than a predetermined ratio value; and

a computer readable medium that stores said computer codes.

3. A computer program product that identifies an unknown base in a sample nucleic acid sequence, comprising:

computer code that receives a first set of signals corresponding to a first set of probe intensities, each probe intensity in said first set indicating an extent of hybridization of a nucleic acid probe with a reference nucleic acid sequence, and each nucleic acid probe differing from each other by at least a single base;

computer code that receives a second set of signals corresponding to a second set of probe intensities, each probe intensity in said second set indicating an extent of hybridization of a nucleic acid probe with said sample sequence, and each nucleic acid probe differing from each other by at least a single base;

computer code that performs a comparison of at least one of said probe intensities in said first set and at least one of said probe intensities in said second set;

computer code that generates a base call identifying said unknown base according to results of said comparisons said sequence of said nucleic acid probe; and

a computer readable medium that stores said computer codes.

4. A computer program product that identifies an unknown base in a sample nucleic acid sequence, comprising:

computer code that receives signals corresponding to statistics about a plurality of experiments, each of said experiments producing probe intensities, each probe intensity indicating an extent of hybridization of a nucleic acid probe with a reference nucleic acid sequence, and each nucleic acid probe differing from each other by at least a single base;

computer code that receives a plurality of signals corresponding to probe intensities, each probe intensity indicating an extent of hybridization of a nucleic acid probe with said sample sequence, and each nucleic acid probe differing from each other by at least a single base;

computer code that performs a comparison of at least one of said plurality of probe intensities with said statistics;

computer code that generates a base call identifying said unknown base according to results of said comparison and said sequence of said nucleic acid probe; and

a computer readable medium that stores said computer codes.

5. A system that identifies an unknown base in a sample nucleic acid sequence, comprising:

a processor; and

a computer readable medium coupled to said processor for storing a computer program comprising:

computer code that receives a plurality of signals corresponding to probe intensities for a plurality of nucleic acid probes, each probe intensity indicating an extent of hybridization of a nucleic acid probe with at least one nucleic acid sequence including said sample sequence, and each nucleic acid probe differing from each other by at least a single base;

computer code that performs a comparison of said plurality of probe intensities to each other; and

computer code that generates a base call identifying said unknown base according to results of said comparison and said sequences of said nucleic acid probes.

6. A system that identifies an unknown base in a sample nucleic acid sequence, comprising:

a processor; and

a computer readable medium coupled to said processor for storing a computer program comprising:

computer code that receives a plurality of signals corresponding to probe intensities for a plurality of nucleic acid probes, each probe intensity indicating an extent of hybridization of a nucleic acid probe with said sample sequence, and each nucleic acid probe differing from each other by at least a single base;

computer code that calculates a ratio of a higher probe intensity to a lower probe intensity; and

computer code that generates a base call identifying said unknown base according to a base of a nucleic acid probe having said higher probe intensity if said ratio is greater than a predetermined ratio value.

7. A system that identifies an unknown base in a sample nucleic acid sequence, comprising:

a processor; and

a computer readable medium coupled to said processor for storing a computer program comprising:

computer code that receives a first set of signals corresponding to probe intensities, each probe intensity in said first set indicating an extent of hybridization of a nucleic acid probe with a reference nucleic acid sequence, and each nucleic acid probe differing from each other by at least a single base;

computer code that receives a second set of signals corresponding to probe intensities, each probe intensity in said second set indicating an extent of hybridization of a nucleic acid probe with said sample sequence, and each nucleic acid probe differing from each other by at least a single base;

computer code that performs a comparison of at least one of said probe intensities in said first set and at least one of said probe intensities in said second set; and

computer code that generates a base call identifying said unknown base according to results of said comparison and said sequence of nucleic acid probe.

8. A system that identifies an unknown base in a sample nucleic acid sequence, comprising:

a processor; and

a computer readable medium coupled to said processor for storing a computer program comprising:

computer code that receives signals corresponding to statistics about a plurality of experiments, each of said experiments producing probe intensities, each probe intensity indicating an extent of hybridization of a nucleic acid probe with a reference nucleic acid

sequence, and each nucleic acid probe differing from each other by at least a single base;

computer code that receives a plurality of signals corresponding to probe intensities, each probe intensity indicating an extent of hybridization of a nucleic acid probe with said sample sequence, and each nucleic acid probe differing from each other by at least a single base;

computer code that performs a comparison of at least one of said plurality of probe intensities with said statistics; and

computer code that generates a base call identifying said unknown base according to results of said comparison and said sequence of said nucleic acid probe.

9. A system according to claims **5**, **6**, **7**, or **8**, wherein the plurality of nucleic acid probes are in an array of probes.

10. A system according to claims **5**, **6**, **7**, or **8**, wherein the plurality of probe intensities are fluorescent intensities.

* * * * *